DOCUMENT RESUME

ED 480 041 CG 032 614

AUTHOR	Helms, Janet E.
TITLE	Fair and Valid Use of Educational Testing in Grades K-12.
PUB DATE	2003-08-00
NOTE	10p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Testing; Elementary Secondary Education; Evaluation Methods; Scoring; *Student Evaluation; *Test Bias; *Test Interpretation; *Test Validity

ABSTRACT

In the United States, standardized educational tests have been used for assessment purposes in grades K through 12 almost since the inception of the testing movement in the early 1900s. Because test-based assessment can have wide-ranging positive or negative effects on K-12 students, the test user must ensure that the tests used for assessment purposes are used fairly and yield valid scores for each student. Fair and valid use of educational testing is most problematic when the student being evaluated differs from the test developer's validation (i.e., norm) group on critical dimensions that might affect the student's responses and reactions to the testing situation or the test user's interpretations of the student's test results. This chapter discusses some of the issues related to fair and valid use of testing for assessment purposes when construct-irrelevant variance is a potential influence on the quality of students' test performance. (GCP)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 041

Fair and Valid Use of Educational Testing in Grades K-12

By
Janet E. Helms

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

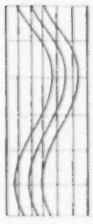
☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

CG032614



Chapter 6
Fair and Valid Use of Educational Testing in Grades K-12

Janet E. Helms

In the United States, standardized educational tests have been used for assessment purposes (e.g., classification and diagnosis) in grades K through 12 almost since the inception of the testing movement in the early 1900s (Domino, 2000).

Assessment refers in part to the process of using test scores to make decisions that affect the educational conditions of individual students. Although the assessment process may involve making use of information obtained from the testing process (e.g., test development, administration, scoring, and interpretation), its focus is the individual rather than the group.

Test scores are used for assessment purposes in the following situations: (a) determining whether a student needs to be placed in a remedial or an accelerated educational environment, (b) permitting a student to advance to the next grade or to graduate, and (c) evaluating the student's mastery of academic content or skills. Because test-based assessment can have wide-ranging positive or negative effects on K–12 students, the test user must ensure that the tests used for assessment purposes are used fairly and yield valid scores for each student.

Fair and valid use of educational testing is most problematic when the student being evaluated differs from the test developer's validation (i.e., norm) group on critical dimensions (e.g., ethnicity, social class, racial socialization, physical abilities) that might affect the student's responses and reactions to the testing situation or the test user's interpretations of the student's test results. On a national level, the K–12 population is characterized by children and adolescents whose home environments reflect a diversity of spoken languages, ethnic and cultural customs and traditions, economic resources, and racial socialization experiences (Helms, 1997). Any of these factors might result in individual test scores measuring constructs that are irrelevant for the intended use of the test. Fair and valid use of tests requires recognition of such construct-irrelevant factors and compensatory efforts to exclude

these factors from the assessment process.

In this chapter, I discuss some of the issues related to fair and valid use of testing for assessment purposes when construct-irrelevant variance is a potential influence on the quality of students' test performance. Although the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) addresses issues of validity and fair testing throughout, chapters 1 ("Validity"), 7 ("Fairness in Testing"), 9 ("Testing Individuals of Diverse Linguistic Backgrounds"), and 10 ("Testing Individuals With Disabilities") are the focus of this chapter.

Valid Use of Testing

Validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). In other words, if the test developer intends a test to be used for particular assessment purposes (e.g., diagnosis, classification), then the test developer must provide the theoretical rationale as to why such usage is appropriate, as well as empirical information that supports such usage. The test user, in turn, must determine whether the focus of the test seems to match her or his assessment needs.

Empirical validity evidence may be obtained in a variety of ways, including correlations between test scores and intended criteria, criterion-group comparisons, and psychometric investigations of the internal structure of the test. Validation methods typically occur at the group level. Validation studies will ordinarily help clarify what cognitive abilities or skills the test under consideration actually seems to measure. It is on the basis of this group-level validity evidence that the test user or educational assessor must make an initial decision with respect to the appropriateness or validity of the test for assessing the individual student.

Decision making on the part of the test user requires specification of the types of criteria that are relevant to the assessment process (i.e., that the test is intended to describe or predict). The assessor should have in mind multiple (ideally nontest) measures of the relevant construct. So, for example, if test scores are being used to assess academic achievement in a particular domain, then alternatives to test scores might be grades or teacher evaluations in that domain. One can have greater faith in the validity of the assessment process when multiple sources present the same picture of the test taker.

A test may be inappropriate for making decisions with respect to a particular student even though validity evidence suggests that the test may be validly used for the typical student. The testing standards or guidelines of most professional assessment organizations advise that test developers describe relevant background characteristics of their norming population as well as the characteristics of the intended test takers (JCTP, 2002). This type of descriptive information should be compared with the characteristics of the student who will be assessed as a means of determining whether there are any obvious differences in background between the student and the test-development sample or the test developer's population specifications. Such discrepancies potentially make the testing process meaningless (i.e., invalid) for the assessed student.

With respect to cultural, racial, physical ability status, and socioeconomic background diversity, the validity of using a test to make decisions about a student from a background different from the test development sample in any of these dimensions may be challenged if the test appears to assess constructs related to background diversity (i.e., construct-irrelevant variance) rather than the construct defined as the stated purpose of the test. For example, if a test written in English is intended to assess students' reading comprehension, but English is a student's second language, then this bilingual student might obtain a low test score because he or she uses the language structures of his or her first language as the model for communicating in English rather than because he or she does not comprehend English text. A test user unfamiliar with this possibility might automatically interpret the low score as a need for remediation in reading skills without examining additional criteria.

When students' irrelevant background information (e.g., social class) influences their test scores, this unintended outcome of the testing process is a source of systematic variance that is irrelevant to assessment of the intended construct (e.g., students' mastery of a mandated curriculum; Helms, 1997). When the test user or assessor has reason to believe that measurement of construct-irrelevant variance in the testing process may have artificially depressed or enhanced a student's performance, he or she should seek confirmation of this hypothesis by examining the a priori alternative criteria. Multiple administrations of the problematic test, however, do not constitute alternative criteria because if assessment of irrelevant constructs is problematic on the first testing occasion, it is likely to be problematic on subsequent testing occasions for the same reasons.

Fair Use of Testing

Whereas validity generally refers to characteristics of the testing process, fair use of tests ultimately refers to the quality of outcomes or decisions resulting from the testing and assessment processes. In general, fairness with respect to testing can be defined as impartial use of tests and interpretation of test results. The current *Standards* (AERA et al., 1999) applies the term *fairness* in the following four ways: (a) tests that are free from bias, (b) equitable treatment of test takers, (c) equality of testing outcomes, and (d) equal opportunity to learn. It might be useful to consider briefly each of these conceptualizations of fair assessment. For an extended examination beyond what I can present here, I refer the reader to specific standards by number (shown in parentheses) as appropriate.

Bias-Free Tests

Deficiencies in a test itself or in the manner in which the test is used in combination with atypical test taker characteristics may result in test scores that differ in meaning across groups of test takers as well as for individual test takers. The existence of bias or lack of bias is ordinarily inferred from comparisons across demographic (e.g., racial or ethnic) groups of the internal structure of tests (e.g., test takers' differential responses to items) or validity evidence. Demographic groups or categories are usually defined according to societal custom and are crude proxies for test-relevant psychological processes (e.g., different response styles) or socialization experiences (e.g., exposure to tested material). Consequently, these demographic categories can be used to describe differences between groups, but not to explain them. If demographic groups differ, the test user must still be able to search for likely explanations of the differences in the student's familial and school socialization experiences.

Differences between groups in average test scores do not necessarily signal the presence of demographic test bias. If empirical studies demonstrate differences in demographic group responses to test content, in response processes used to answer test items, or in empirical validity evidence, then the test developer should collect separate validity data for the counter-normative as well as the normative examinee population (Standards 7.1, 7.2, 7.6, 7.11, 9.2). Moreover, to rule out demographic group bias, local test users should collect validity information in their own settings to make sure that test scores are not

misrepresenting the abilities, knowledge, or skills of the affected students—particularly if the local student population is known to differ from the larger examinee population with respect to demographic background characteristics.

Haney (1993) reported that the College Board (an affiliation of 2,500 schools and colleges) has offered to help colleges perform local validity studies. Presumably, assessors who use tests to make decisions about students in grades K through 12 could also require such services from test developers as a condition for using their tests. Nevertheless, data relevant to demographic test bias as it pertains to individual students may not be available to aid the assessor in interpreting students' test scores. In such cases, common sense will have to prevail. If bias cannot be ruled out as a factor, then the test user should consider the appropriateness of using within-group scoring criteria (e.g., local cutoff scores) as the basis of assessment decisions.

Equitable Treatment

The concept of equitable treatment in the testing process means impartial treatment at every phase of the process. All test takers should be tested under equivalent as opposed to the same testing conditions. For example, unless the stated or intended rationale for test use is assessment of proficiency in the language of the test, then all test takers should have the opportunity to be tested in the language in which they are most proficient (Standards 9.3, 9.4). Test developers may include in their test manuals information about appropriate test accommodations for ensuring equivalence of testing conditions with respect to various demographic groups (Standards 10.1, 10.4), but in case they do not, test users should familiarize themselves with available empirical information as well as relevant testing law to help inform their assessment decision making.

Equitable treatment also involves ensuring that test takers have comparable opportunities to become familiar with the structure of the testing process. A fair testing structure includes appropriate testing conditions and equal opportunities for test takers to familiarize themselves with the test format, practice materials, and related material properties of the testing situation that might be expected to interfere unfairly with a student's test performance. Moreover, if the test user is aware that the student's performance may be enhanced by special preparation routinely available to other students (e.g., coaching), then the test taker or the test taker's guardian should be so advised.

Equality in Testing Outcomes

In the testing literature, fairness of testing outcomes generally refers to whether the use of test scores unfairly penalizes demographic group members with respect to selected outcomes (e.g., selection, promotion, or graduation). As previously mentioned, differences between groups in test-based outcomes do not necessarily mean that the testing process is biased against certain groups or individual members of such groups. Limitations in testing methodology (e.g., less-than-perfect correlations between criteria and test scores), however, make it impossible to rule out test bias or unfair use of tests as possible explanations for between-group differences in outcomes. Common practice among testing professionals is to use such observed differences as inspiration for further study of the tests or to infer fairness from relevant validity evidence, assuming that such evidence has been obtained under equitable testing conditions for all groups.

Although some professional testing standards require that test users and developers remove test score variance that is unrelated to the skills or abilities that are the focus of the assessment, objective techniques for doing so are not commonly used (Helms, in press). If test users or assessors can identify appropriate outcome-relevant validity evidence, they may use inductive reasoning to form hypotheses about whether outcome decisions affecting individual students from atypical backgrounds are fair. Multiple criteria related to the intended outcome will be useful for this purpose.

Opportunity to Learn

Fairness also refers to the extent to which test takers have had comparable opportunities to learn the material covered by the test. This use of fairness, which is typically of concern when achievement tests are used as the basis for decision making, is perhaps the most controversial. Fair use of tests with respect to this definition requires that the test user differentiate the test taker's access to specific resources (e.g., tested material) from her or his relevant intellectual skills or abilities.

For example, a student might receive a low score on a mathematics achievement test because the test covered material to which he or she had not been exposed. If the student's grades in mathematics courses suggest superior skills, then the student's low test score might reflect a difference in opportunity rather than a lack of relevant skills. In such situations, the testing process has assessed construct-irrelevant variance (i.e., deficient curriculum content). Consequently, assessment decisions

that penalize the student (e.g., grade retention) are unfair under this definition of fairness. The test user has a responsibility to review test content in combination with relevant factors in the test taker's school environment to help prevent this type of unfair use of tests (JTCP, 2002).

Conclusion

As the role of tests in students' lives grows in significance, test users must acknowledge the diversity of the school-age population as an important aspect of test development and test use. Many student characteristics and environmental conditions and practices may interact and contribute to systematic variance that is irrelevant to measurement of the construct of interest to the test user or assessor. Fair and valid use of tests for making high-stakes decisions affecting children and adolescents requires attention to the racial, cultural, physical ability, and other background factors that may differentially influence individual students' performance on such measures relative to the comparison groups on which the tests were developed. Moreover, fair and valid use of tests for assessment purposes means that the test user sometimes must base high-stakes decisions on the characteristics of the students and schools in which the student functions rather than on national norms or comparison groups.

References

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Domino, G. (2000). *Psychological testing: An introduction.* Upper Saddle River, NJ: Prentice Hall.

Haney, W. (1993). Testing and minorities. In L. Weis & M. Fine (Eds.), *Beyond silence: Class, race, and gender in United States schools* (pp. 45–73). Albany, NY: State University of New York Press.

Helms, J. E. (1997). The triple quandary of race, culture, and social class. In D. Flanaghan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 517–532). New York: Guilford Press.

Helms, J. E. (in press). A remedy for the Black-White test-score disparity. *American Psychologist.*

✦JCTP [Joint Committee on Testing Practices]. (2002). *Code of fair testing practices in education.* Available on *Measuring Up: An Anthology of Assessment Resources* [CD]. Also retrievable on-line: http://aac.ncat.edu.

✦ Document is included in the Anthology of Assessment Resources CD

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").